DATA SCIENCE WITHOUT BORDERS

WES MCKINNEY @WESMCKINN

JupyterCon | August 2017
The information presented here is offered for informational purposes only and should not be used for any other purpose (including, without limitation, the making of investment decisions). Examples provided herein are for illustrative purposes only and are not necessarily based on actual data. Nothing herein constitutes: an offer to sell or the solicitation of any offer to buy any security or other interest; tax advice; or investment advice. This presentation shall remain the property of Two Sigma Investments, LP (“Two Sigma”) and Two Sigma reserves the right to require the return of this presentation at any time.

Some of the images, logos or other material used herein may be protected by copyright and/or trademark. If so, such copyrights and/or trademarks are most likely owned by the entity that created the material and are used purely for identification and comment as fair use under international copyright and/or trademark laws. Use of such image, copyright or trademark does not imply any association with such organization (or endorsement of such organization) by Two Sigma, nor vice versa.

Copyright © 2017 TWO SIGMA INVESTMENTS, LP. All rights reserved
THINKING ON THE LAST 10 YEARS

2007 → 2017
CLOSED SOURCE  →  OPEN SOURCE
A shared front-end for data science
THE NEXT 10 YEARS AND BEYOND

2017 ➞ 2027 ➞ ...

7
THE AI ARMS RACE
CHANGING HARDWARE LANDSCAPE

DISK

PROCESSING

MEMORY
DATA SCIENCE “LANGUAGE “SILOS”

FRONT-END

PYTHON  R  JVM  JULIA  ...

10
WHAT’S IN A SILO?

- STORAGE / DATA ACCESS
- DATA STRUCTURES / IN-MEMORY FORMATS
- GENERAL COMPUTE ENGINE(S)
- ADVANCED ANALYTICS
WHAT’S IN A SILO?

STORAGE / DATA ACCESS

DATA STRUCTURES / IN-MEMORY FORMATS

GENERAL COMPUTE ENGINE(S)

ADVANCED ANALYTICS

pandas

NumPy pandas

NumPy pandas

scikit-learn
RENOVATING PANDAS
MAKING THE SILOS “SMALLER”

FRONT-END

PYTHON  R  JVM  JULIA  ...

?
PROGRAMMING LANGUAGES AS USER INTERFACES
SAME ANALYSIS, DIFFERENT IMPLEMENTATION

R

```r
df <- read_csv(…)
df % group_by(…) % summarise(…)
```

PYTHON

```python
df = read_csv(…)
df.groupby(…).aggregate(…)
```
A SHARED RUNTIME FOR DATA SCIENCE

FRONT-END

PYTHON  R  JVM  JULIA  ...

SHARED DATA SCIENCE RUNTIME
FROM IDEA TO ACTION
PART 1: STANDARD IN-MEMORY FORMAT

Non-Portable Data Frames

PYTHON
R
JVM

PORTABLE DATA FRAME
PART 2: ZERO COPY INTERCHANGE

PYTHON  R  JVM ...

SHARED MEMORY + STANDARD MEMORY FORMATS
PART 3: HIGH PERFORMANCE DATA ACCESS

Storage Formats/ Databases

- CSV
- BINARY COLUMNAR
- SQL

PORTABLE DATA FRAME
PART 4: FLEXIBLE COMPUTATION ENGINE

- Zero-overhead User-defined Functions
- Portable Operator “Graphs”
- “Embeddable” in Larger Systems
APACHE ARROW
Language-agnostic Data Frame Format
Zero-Copy Interchange
BUILDING THE ARROW FORMAT

• “Superset” of representations supported by R, pandas, SQL engines

• Optimized for CPU cache affinity

• ASF Governance: Open + Transparent Community Project
FEATHER: MINIMALIST ARROW ON DISK
Some Arrow OSS Users

- goai
- Ray Project
- dremio
- Apache Spark
- Turbodbc
- Geomesa
- pandas
- Feather Format
- Parquet
- DASK
BUILDING THE FUTURE
THANK YOU

WES MCKINNEY @WESMCKINN
Apache Arrow: http://arrow.apache.org